


RESEARCH ARTICLE

Open Access



Developing an efficient DNA barcoding system to differentiate between *Lilium* species

Yixin Liu¹, Mingfang Zhang¹, Xuqing Chen¹, Xi Chen^{1,3}, Yue Hu¹, Junlian Gao¹, Wenqiang Pan², Yin Xin^{1,2}, Jian Wu^{2*} , Yunpeng Du^{1*} and Xiuhai Zhang^{1*}

Abstract

Background: *Lilium* is an important ornamental bulb, possesses medicinal properties, and is also edible. Species within the *Lilium* genus share very similar morphology and macroscopic characteristics, thus they cannot be easily and clearly distinguished from one another. To date, no efficient species-specific markers have been developed for classifying wild lily species, which poses an issue with further characterizing its medicinal properties.

Results: To develop a simple and reliable identification system for *Lilium*, 45 representative species from 6 sections were used to develop a DNA barcoding system, which was based on DNA sequence polymorphisms. In this study, we assessed five commonly used DNA barcode candidates (*ITS*, *rbcl*, *ycf1b*, *matK* and *psbA-trnH*) and five novel barcode candidates obtained from highly variable chloroplast genomic regions (*trnL-trnF*, *trnS-trnG*, *trnF-ndhJ*, *trnP-psaJ-rpl33* and *psbB-psbH*). We showed that a set of three novel DNA barcodes (*ITS* + *trnP-psaJ-rpl33* + *psbB-psbH*) could be efficiently used as a genetic marker to distinguish between lily species, as assessed by methods including DNAsp, BI and ML tree, and Pair Wise Group (PWG).

Conclusions: A rapid and reliable DNA barcoding method was developed for all 45 wild *Lilium* species by using *ITS*, *trnP-psaJ-rpl33*, and *psbB-psbH* as DNA barcoding markers. The method can be used in the classification of wild *Lilium* species, especially endangered species, and also provides an effective method for selective lily breeding.

Keywords: *Lilium*, cpDNA hypervariable region, DNA barcoding, Species identification, Phylogenesis

Background

Lilium is a genus of flowering plants which includes 110–115 species of lily. Lilies are endemic to the cold and temperate regions of the northern hemisphere [1–4]. Lilies have originated from the Himalayas and the southwest of China, where there are currently about 51 species/varieties present [5, 6]. Based on their morphological

classification, these species are divided into the following 8 sections: Section *Martagon*, Sect. *Pseudolirium*, Sect. *Liriotypus*, Sect. *Archelirion*, Sect. *Sinomartagon*, Sect. *Leucolirion*, Sect. *Lophophorum*, and Sect. *Lilium*—*Nomocharis* [7]. Today, more than 10,000 cultivars have been registered in the Royal Horticulture Society, showcasing great value around the world [8]. Lily cultivars are also classified into 8 groups (Oriental hybrid, LA-hybrid, OT-hybrid, Asiatic hybrid, LO-hybrid, Longiflorum, and Aurelian & Trumpet), which are cultivated to create ornamental value [2, 4]. *Lilium* species are not only used as ornamental plants but also to produce food and medicine, especially in Asia. The traditional edible lilies mainly include *L. davidii* var. *willmottiae*, *L. brownii*, *L. lancifolium*, *L. longiflorum* [9]. Moreover, *L. lancifolium*, *L. pumilum* and *L. brownii* are used as a component of

*Correspondence: jianwu@cau.edu.cn; dyp_851212@126.com; zhangxiuhai@baafs.net.cn

¹ Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China

² Beijing Key Laboratory of Development and Quality Control of Ornamental Crops, Department of Ornamental Horticulture and Landscape Architecture, China Agricultural University, Beijing 100193, China

Full list of author information is available at the end of the article



© The Author(s) 2021. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

traditional Chinese medicine for lung ailments [10, 11]. Due to their ornamental, edible, and medicinal properties, *Lilium* species have great commercial value [9].

However, it is difficult to clearly identify *Lilium* species within the same category solely based on their highly similar morphological characteristics [12–14]. Additionally, as a result of overharvesting and reduction in their natural habitat, the abundance of wild lily species has dramatically decreased [9, 15, 16]. This has led to several species being listed as key protected wild plants, including *L. fargesii*, *L. amoenum*, *L. henrici*, *L. paradoxum*, *L. taliense*, and *L. wardii* [17–20]. Hence, the ability to correctly identify endangered lily species using genetics may serve a role in effectively protecting these species. Furthermore, the high species diversity found within the same ecosystem further muddles our process to correctly identify lily species based on traditional morphology-based taxonomy, resulting in ambiguous phylogenetic classification. DNA marker technology, such as the use of DNA barcoding in classification systems, has significantly enhanced the identification, protection, and sustainable use of plant resources. However, no high-resolution markers for *Lilium* have been developed yet.

DNA barcoding is a system that involves the sequencing of short DNA fragments for fast and accurate species identification, especially suitable for highly homologous species [21, 22]. DNA barcoding is a widely used method for distinguishing macroscopically similar species, detecting the spatial distribution of plant roots, and studying invasive plant species [22]. Moreover, the genetic sequences used in DNA barcoding can be conserved enough across species to facilitate the design of universal sequencing primers. DNA barcoding is also beneficial to the fields of conservation and evolutionary ecology [23], where this method is utilized to evaluate the genetic diversity of endangered species needed for their protection and population restoration [24]. Universal barcodes, such as *rbcL*, *matK*, and *psbA-trnH*, use sequences from chloroplast genes and have been previously identified in different species [25–30]. However, apart from universal barcodes, we also require the development of species-specific DNA barcodes [31–33].

With the rapid development of next-generation sequencing, obtaining chloroplast-specific genomic sequences has become much easier and can be used in extending gene-based phylogenetics to phylogenomics [34, 35]. Differences in chloroplast (cp) genomic sequences between plant species can be comprehensively applied in the phylogenetic classification of individual plants [36]. While the cp genome contains highly conserved regions, the highly variable genomic regions could be used to identify DNA barcoding candidates which

would resolve the phylogenetic relationship between species.

Recently, molecular phylogenetic studies in *Lilium* were carried out using molecular markers like RAPD (Random Amplified Polymorphic DNA) and ISSR (Inter-simple Sequence Repeat), which have low rates of PCR amplification and low sequencing resolution [39, 40]. In ideal conditions, a barcode should be variable enough to resolve closely related species and short enough for easy experimental manipulation with low cost. Therefore, in this study, we investigated the effectiveness of using DNA barcoding to distinguish between 45 representative *Lilium* interspecies and developed an efficient DNA barcoding system by using a combination of genetic markers (*ITS*, *trnP-psaJ-rpl33*, and *psbB-psbH*).

Results

Hypervariable region assessment within the chloroplast genome in *Lilium* spp.

To identify potential DNA barcodes, we screened previously published genomic data from 16 *Lilium* chloroplast (cp) genomes for hypervariable regions [19]. By using multiple sequence alignment, we identified the following hypervariable regions: *trnS-trnG*, *trnE-trnT-psbD*, *trnF-ndhJ*, *psbE-petL*, *trnP-psaJ-rpl33*, *psbB-psbH*, *petD-rpoA*, *ndhF-rpl32-trnL*, *ycf1a*, and *ycf1b* [19]. A total of 521 nucleotide polymorphisms were identified in these 10 regions, which may be used as genetic markers for phylogenetic construction and species identification in *Lilium*. In order to evaluate the nucleotide polymorphism information (Pi), the stepwise genetic nucleotide diversity (π ; Nei 1987) was estimated within the 10 regions. The π , representing the existence of different nucleotide bases between species, ranged from 0.01364 to 0.01833 within the 16 species. This data indicates the presence of mutations in the hypervariable regions of the relatively stable chloroplast genome, which could be used to develop candidate DNA barcodes [19].

We then successfully amplified five of the ten hypervariable regions by PCR. To develop high-resolution DNA barcoding, we added another five canonical plant barcoding markers (Table 2). Among the candidate barcodes, *psbA-trnH* was the shortest in length (400 bp) and *psbB-psbH* was the longest (1124 bp), with an average size of 727 bp. Ten DNA barcodes were then used to amplify sequences from 45 species used, plus an additional three outgroups (*Cardiocrinum giganteum*, *Nomocharis pardanthina*, and *Fritillaria karelinii*). This resulted in the amplification of 472 different sequences. The sequencing success rate of *trnF-ndhJ* (93.7%) and *ITS* (91.6%) were relatively low, while the rest of the DNA barcodes had 100% sequencing success rates (Table 4).

Barcoding assessment using chloroplast and nuclear markers in *Lilium* spp.

To assess the efficiency of the DNA barcodes identified, we evaluated eight different barcoding indices to assess the potential for each candidate sequence to be useful in lily DNA barcoding. The total number of mutations (Eta), haplotype diversity (Hd), π , G + C ratio, information site, and average nucleotide difference (k), were analyzed for all 472 amplified DNA sequences (Table 3). The Hd index was highest for the following genomic sequences: *ITS* (0.995), *trnF-ndhJ* (0.965), *ycf1b* (0.952), *matK* (0.94), and *psbB-psbH* (0.926). This data shows that these genomic regions would have the highest allelic difference in randomly sampled individuals and could thus serve as a useful parameter in distinguishing individual species. We then selected those five regions for further analysis to determine their potential as DNA barcodes. The genetic nucleotide diversity (π) was high, ranging from 0.03035 to 0.30907. The total number of mutations (Eta) of the five candidate barcodes was also higher compared to other sequences (244 to 372, except for *trnF-ndhJ* with an Eta of 49). Notably, the lowest overall scores were mainly given by the analysis of the conserved DNA barcodes. For example, *psbA-trnH* had a Hd score of 0.792. Taken together, all the above analyses show that we identified five possible barcoding regions: *trnP-psaJ-rpl33*, *psbB-psbH*, *ycf1b*, *ITS*, and *matK*.

To estimate the genetic diversity between interspecies in *Lilium*, these five candidate barcodes were chosen for combinatorial barcode analysis by Kimura-2-Parameter (K2P). Overall, 19 combinations of candidate barcodes were obtained, and the results showed that 1) there were no barcode gaps when combining two barcodes (Fig. 1); 2) the highest variation in interspecific divergence resulted from the combination of three barcodes (*ITS + trnP-psaJ-rpl33 + psbB-psbH*) (Fig. 2); and 3) the interspecific differentiation was relatively low, and the diversity could not be clearly observed when combining four or five barcodes (Figs. 3 and 4). These data demonstrated that a combination of three candidate barcoding sequences could be the most efficient in distinguishing between lily species.

Lilium spp. identification using DNA barcoding

To evaluate the barcoding gap in comparing with the distribution of the pair-wise interspecific distance for single/ combined barcode(s), we analyzed DNA barcoding sequences using Taxon DNA. The 'Best match' analysis was then performed to determine the closest barcode match for given sequences, regardless of the sequence similarity in the barcoding sequence. This meant that every queried sequence would be assigned

the best matching barcode. When comparing the results of the 'Best match' analysis and 'Best close match' analysis, the former always presented higher or equal individual identification rates (Table 4). The barcode index of combined candidates was generally higher than that of a single candidate barcode. This analysis demonstrates that the barcoding combination of *ITS + trnP-psaJ-rpl33 + psbB-psbH* has the highest potential success rate (12.33%) in correctly identifying the lily species, followed by *ITS+trnP-psaJ-rpl33* (10.41%).

Tree-based analysis with chloroplast and nuclear DNA barcoding in *Lilium* spp.

To validate the resolution for using *ITS+trnP-psaJ-rpl33+psbB-psbH* as DNA barcoding to identify *Lilium* species, we verified that the candidate barcodes can clearly distinguish between species by constructing a phylogenetic tree. We used 45 representative species with three biological replicates each, by using maximum likelihood (ML) and Bayesian (BI) Phylogenetic tree. The results from both the ML and BI tree analyses divided these species into four sections and eight groups (Fig. 5). The four sections were Sect. *Sinomartagon*, Sect. *Archelirion*, Sect. *Martagon*, and Sect. *Leucolirion*. Within the Sect. *Sinomartagon*, the resulting identity was 96%, so this category was further divided into four groups: Group 1, Group 3, Group 6, and Group 8. A similar analysis was carried out within the Sect. *Archelirion*, where the species were divided into two groups, Group 2 and Group 7. Sect. *Leucolirion* only had one group (Group 5) and Sect. *Archelirion* mainly belonged to Group 4. The discrimination rate in Group 8 was 86%. Additionally, *Nomocharis pardanthina* and *L. lophophorum* in Group 8 clustered together in the same branch with an approval rating of [BI=1, ML=94], indicating that *N. pardanthina* was genetically closer to *Lilium*. Overall, the data validated the use of the newly identified DNA barcodes in accurately distinguishing species of *Lilium*.

Discussion

Lilium is one of the most valuable ornamental plant genera worldwide, with 100–115 species and over 10,000 cultivars. However, until now, an efficient method for distinguishing *Lilium* spp. had yet to be established. Here, we developed and validated a set of high-resolution DNA barcodes for use as a tool in distinguishing 45 species representing six sections.

Discrimination comparison among different analysis methods

Previously, several methods were generally used for analyzing interspecies discrimination with DNA barcoding, including phylogenetic trees (NJ, BI, MP, ML and

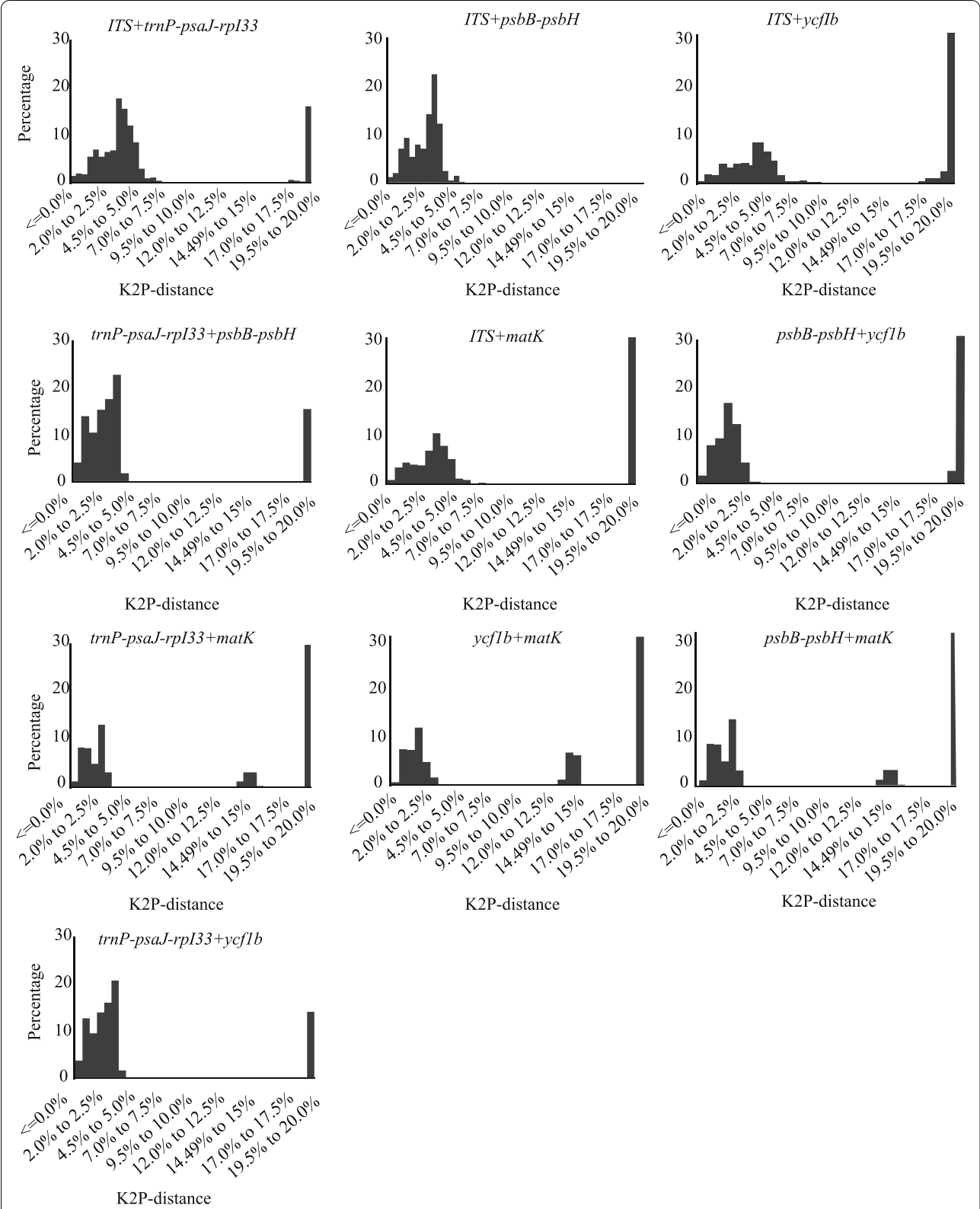
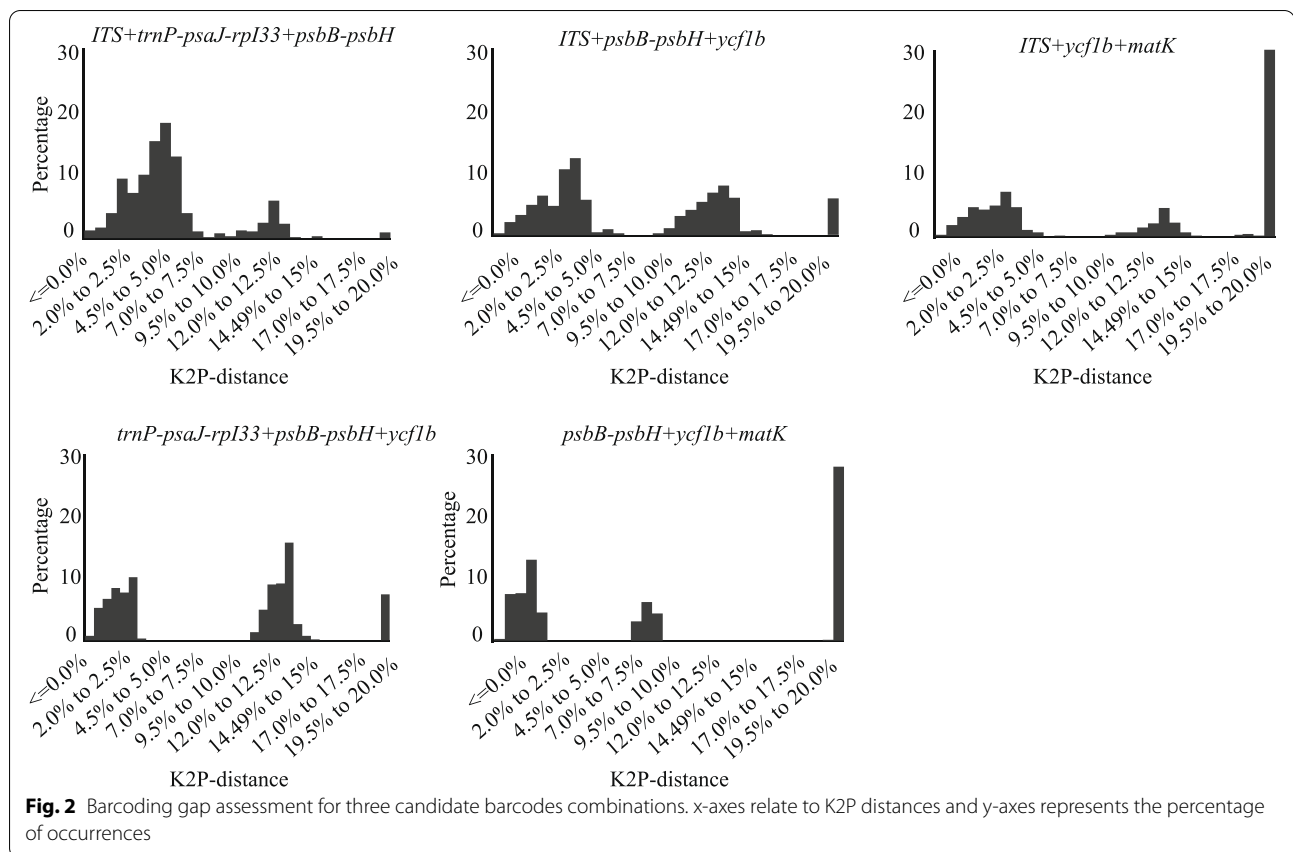


Fig. 1 Barcoding gap assessment for two candidate barcode combinations. x-axes relate to K2P distances and y-axes represent the percentage of occurrences



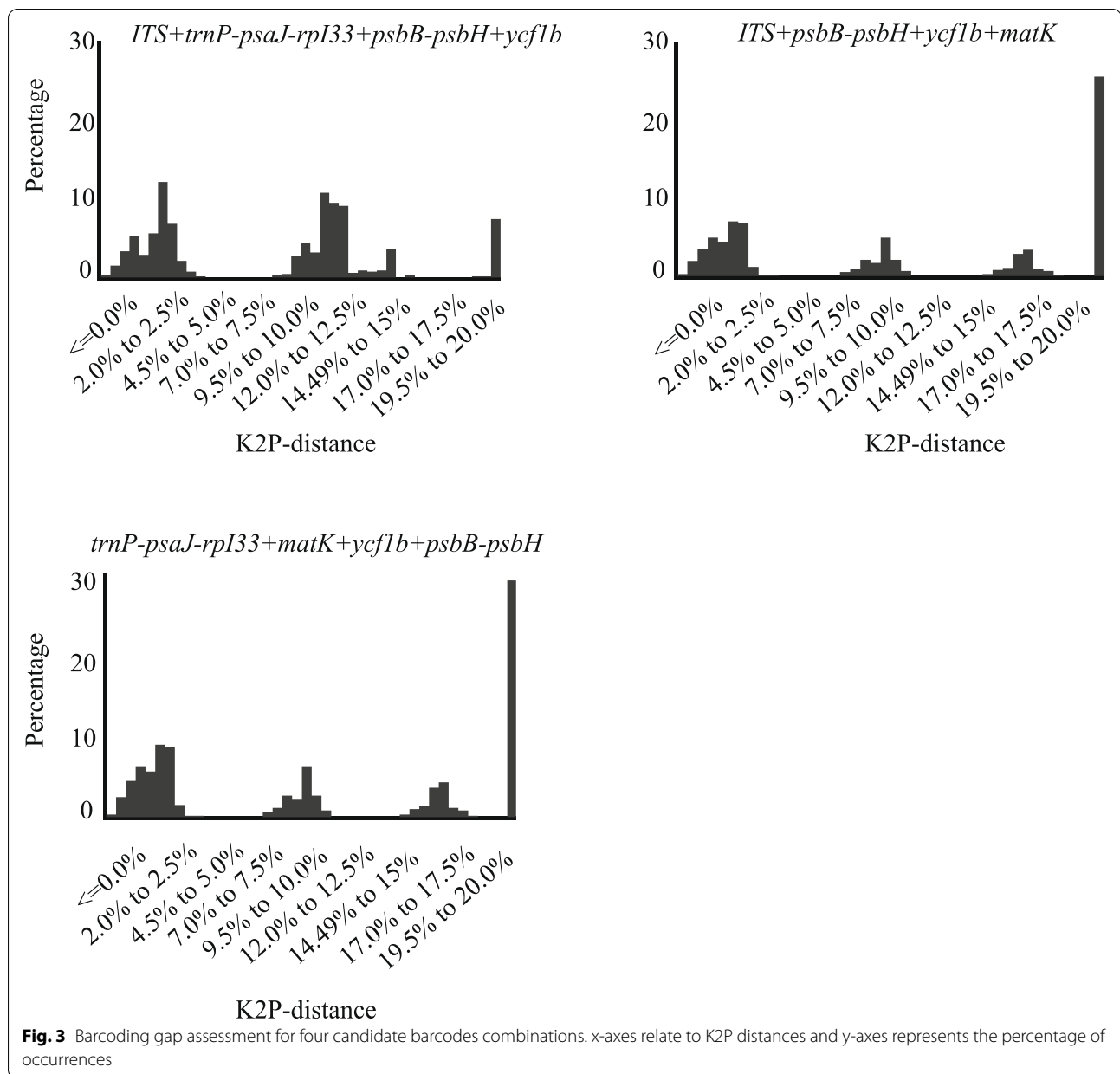
et al.), distance-based (PWG, P-distance, K2P distance, and et al.), and character-based methods (BLOG, DNA-BAR, BRONX, and et al.) [41, 42]. However, there was no standard method in place for species identification [43, 44]. In this study, we used various methods to develop DNA barcoding specifically for species within the *Lilium* genus.

First, we used Taxon DNA analysis (simple pairwise matching for DNA barcoding) to evaluate the *Lilium*. The index from 'Best match' in Taxon DNA analysis was lower than those from the tree-based analysis. Thus, tree-based analysis was more suitable in determining DNA barcoding sequences for identifying different lily species. In Taxon DNA analysis, small barcoding gaps, high similarity in interspecies, unclear origin and obscure evolution contributed to low discrimination [19]. Next, we verified the candidate DNA barcodes by using BI and ML methods to construct the phylogenetic tree. To find a more accurate model, ML was repeated one thousand times and BI tree was repeated one billion times, which generated a more definite analysis than an NJ tree [45]. Then, we performed phylogenetic tree analysis to evaluate the identified candidate DNA barcodes in *Lilium* and revealed that Sect. *Sinomartagon* and Sect. *Archelirion*

were more differentiated than the other sections. In the Group 8, *N. aperta* was close to *N. pardanthina* [46] but nested within *Lilium*, which was clustered in the same branch as the *L. duchartrei* of Sect. *Sinomartagon* clustered. Overall, the genetic analysis of DNA barcodes gave similar classification results compared to morphological classification, which demonstrates the reliability of the recommended barcodes. However, using DNA barcodes to identify between species is a superior classification method to using morphological characteristics. Using DNA barcoding for successful identification at the subgenus or node level could also be considered if individuals form a monophyletic clade [19]. Additionally, the appropriate phylogenetic methodology could provide a reliable reference for the study of the origin, phylogenetic, and differentiation ages, and help solve the phylogenetic relationship and classification complications in *Lilium*.

The Evaluation of *Lilium* DNA Barcode

In this study, hypervariable regions of *Lilium* chloroplast genome and conserved plant fragments were selected as candidate DNA barcodes. An ideal DNA barcode requires clear species discrimination and high-quality primer pairs [47, 48]. According to the criteria above,



we used eight indices (Table 3) to evaluate the candidate barcode index and five of those fifteen barcodes showed the highest indices quality and bidirectional sequences. The five identified candidate DNA barcodes were *matK*, *trnP-psaJ-rpl33*, *psbB-psbH*, *ycf1b*, and *ITS*. *ITS* is the only DNA sequence that belongs to a conserved ribosomal DNA genetic region, while the others belong to the hypervariable regions of cp genome DNA in the plant. This indicated that the mutation information in the hypervariable regions of the relatively stable chloroplast genome was suitable for developing candidate DNA barcodes for *Lilium spp.* identification.

As previous studies have reported, due to its high level of sequencing success rate (91.6%), the haplotype diversity (0.995), and the number of haplotypes (40), *ITS* provided the highest species resolution. *ITS* also has the highest species resolution as measured by “Best match (9.09%)”. Therefore, *ITS* is regarded as a suitable candidate for plant DNA barcoding and has been widely used in community phylogeny and biodiversity surveys [49]. We obtained similar results in *Lilium*, where *ITS* barcoding indices were significantly higher than those of other candidate DNA barcodes (Table 3, Table 4). To our excitement, we found that *trnP-psaJ-rpl33* and *psbB-psbH*

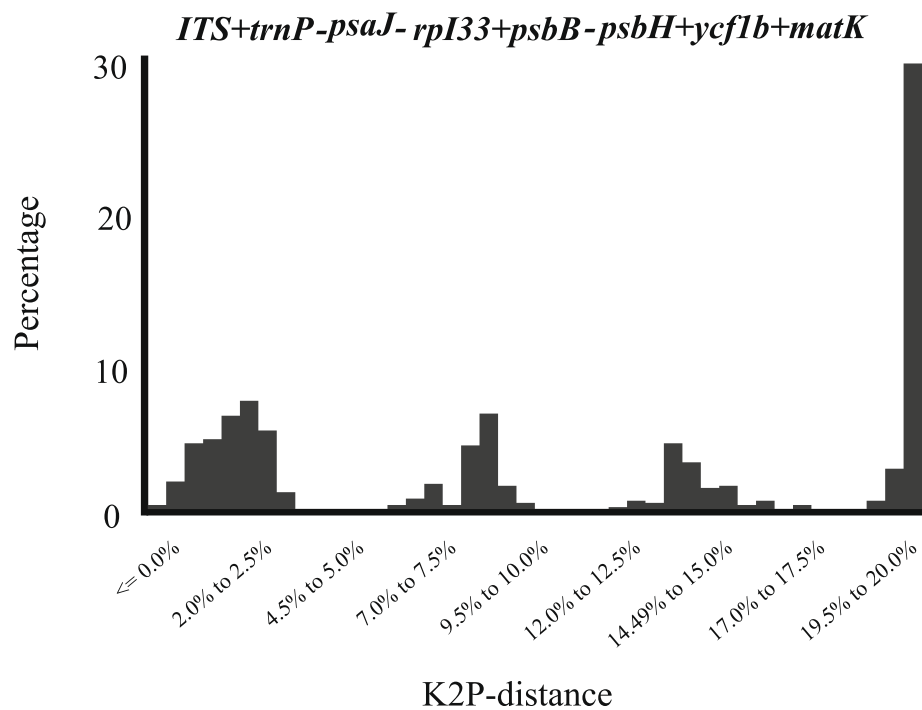


Fig. 4 Barcoding gap assessment for five candidate barcodes combinations. x-axes relate to K2P distances and y-axes represents the percentage of occurrences

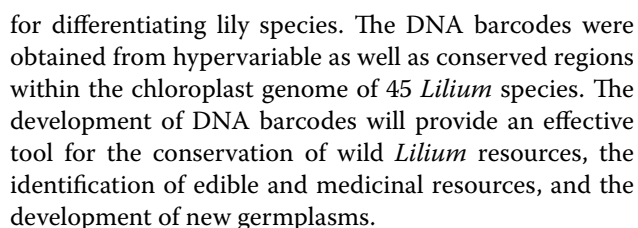
yielded the most promising results as DNA barcodes since this combination had a sequencing success rate of 100% (Table 3). Furthermore, several reports have been previously published regarding the use of *matK* in DNA barcoding, which belongs to the conserved cp genetic regions [21, 50], and has thus led some researchers to have reservations about the use of this locus. In the current study, the *matK* locus was a relatively high information locus (192) and the success rate of sequencing using this locus was 100%. These data made *matK* as one of the candidate DNA barcodes in *Lilium*. We also noticed that the *matK* sequence was easy to amplify but it often performed poorly in complex evolving groups [51], which was consistent with previous reports. Overall, *ITS*, *trnP-psaJ-rpI33*, and *psbB-psbH* were the best single-barcode candidates for *Lilium* identification.

The use of a combination of barcodes can often improve the ability to identify species [52]. Of the 19 barcode combinations tested, *ITS + trnP-psaJ-rpI33 + psbB-psbH* performed best in the 'Best match (12.33%)', 'Best close match (12.33%)' and 'all species barcodes (8.08%)'. The results of the combination of the two barcodes showed (Fig. 1) that combinations containing *ITS* significantly improved the resolution of species identification. *ITS + trnP-psaJ-rpI33* (10.41%) and *ITS + psbB-psbH* (8.33%) had higher discriminatory power than *psbB-psbH + matK* (6.25%) and *ycf1b + matK*

(4.16%). In the analysis of three-barcode combinations (Fig. 2), we found that *ITS+psbB-psbH+ycf1b* (8.33%) has the same "Best Match" index as *ITS+psbB-psbH* (8.33%). Surprisingly, we were not able to increase the resolution success by combining four DNA barcodes. This may be related to the complexity of *Lilium* genome and the low resolution in distinguishing between sequences found within the conserved fragments. Considering discriminatory power, cost-efficiency and effort, the three-marker combination '*ITS+trnP-psaJ-rpI33+psbB-psbH*' showed the best species identification among all the compared marker combinations, suggesting that it may be the best choice for barcoding in *Lilium*. Although various barcodes have been widely used in different plants before, the species classification within a specific genus is affected by many factors, which often results in the uncertainty of a single genetic site. Therefore, in order to develop an appropriate plant DNA barcoding, selected markers should not emphasize universality in all plant species but should be more specific for a certain taxon.

Conclusions

In summary, lily is a highly valuable ornamental and medicinal plant. In this study, by constructing the phylogenetic tree, we identified that a combination of three DNA barcodes was the most effective method



Plant materials

All plant materials are original from the National Lily Germplasm Bank at Beijing Academy of Agricultural and Forestry Sciences. To capture high-resolution genetic diversity, samples were collected in largescale and extensive distribution. Our endemic wild *Lilium* belonging

Table 1 Information about tested *Lilium* species in this study

| Section (Comber) | No. of tested species | Species |
|---------------------------|-----------------------|---|
| Sect. <i>Martagon</i> | 6 | <i>L. martagon</i> , <i>L. distichum</i> , <i>L. tsingtauense</i> , <i>L. hansonii</i> , <i>L. martagon</i> var. <i>pilosiusculum</i> , <i>L. henrici</i> |
| Sect. <i>Archelirion</i> | 5 | <i>L. rubellum</i> , <i>L. speciosum</i> var. <i>gloriosoides</i> , <i>L. auratum</i> , <i>L. japonicum</i> , <i>L. nepalense</i> var. <i>ochraceum</i> |
| Sect. <i>Sinomartagon</i> | 20 | <i>L. amabile</i> , <i>L. davidii</i> , <i>L. cernuum</i> , <i>L. leichtlinii</i> var. <i>maximowiczii</i> , <i>L. leichtlinii</i> , <i>L. dauricum</i> , <i>L. davidii</i> var. <i>willmottiae</i> , <i>L. duchartrei</i> , <i>L. taliense</i> , <i>L. lankongense</i> , <i>L. callosum</i> , <i>L. fargesii</i> , <i>L. lophophorum</i> , <i>L. primulinum</i> var. <i>ochraceum</i> , <i>L. bulbiferum</i> , <i>L. pumilum</i> , <i>L. concolor</i> var. <i>pulchellum</i> , <i>L. lancifolium</i> , <i>L. henryi</i> , <i>L. rosthornii</i> |
| Sect. <i>Leucolirion</i> | 12 | <i>L. sulphureum</i> , <i>L. regale</i> , <i>L. brownii</i> , <i>L. leucanthum</i> , <i>L. leucanthum</i> var. <i>centifolium</i> , <i>L. formosum</i> , <i>L. sargentiae</i> , <i>L. longiflorum</i> , <i>L. bakerianum</i> var. <i>aureum</i> , <i>L. bakerianum</i> var. <i>delavayi</i> , <i>L. bakerianum</i> var. <i>rubrum</i> , <i>L. bakerianum</i> |
| Sect. <i>Pseudolirium</i> | 1 | <i>L. pardalinum</i> |
| Sect. <i>Liriotypus</i> | 1 | <i>L. candidum</i> |

Table 2 Screening results of each barcode primer

| Gene Name | Length (bp) | Primer Name | Forward primer | Reverse primer |
|------------------------|-------------|-------------|--------------------------|-------------------------|
| <i>trnF-ndhJ</i> | 855bp | LHV3 | TGGATATAGACCTCCATTTTGGAG | GATAATGACACGACTCCAGAA |
| <i>trnS-trnG</i> | 665bp | S1 /LHV1 | CTCTCCCAACTCAAATTG | CAGAATTATGAAAATTATAGCGT |
| <i>trnP-psaJ-rpl33</i> | 633bp | S5/LHV5 | ATCCTTGCTTGTTTCCAC | TTCTAACTMTCAATTATTCCTA |
| <i>psbB-psbH</i> | 1124bp | LHV6 | GGGTTGGTTCACCTTTTGGGC | TCCACGGTCGAACTACCAGA |
| <i>ycf1b</i> | 750bp | LHV10 | ACCACCGCTTTGGCTTTTCT | CCATGCCCATTTCGGTTTG |
| <i>matK</i> | 800bp | matK | CGATCTATTCATTCAATATTTTC | TCTAGCACACGAAAGTCGAAGT |
| <i>rbcL</i> | 600bp | rbcL | ATGTCACCACAAACAGAGAC | TCACATGTACCCGAGTAGC |
| <i>psbA-trnH</i> | 400bp | psbA-trnH | ACTGCCTTGATCCACTTGGC | CGAAGCTCCATCTACAAATGG |
| <i>trnL-trnF</i> | 700bp | trnL-trnF | CGAAATCGGTAGACGTACG | ATTTGAAGTGGTGACACGAG |
| <i>ITS</i> | 750bp | ITS | GGAAGKARAAGTCGTAACAAGG | RGTTTCTTTTCTCCGCTTA |

Table 3 Analysis of different barcoding indices of *Lilium*

| | Hd | H | Pi | k | Eta | G+C Ratio (%) | Information site | success rate of sequences |
|------------------------|-------|----|---------|--------|-----|---------------|------------------|---------------------------|
| <i>matK</i> | 0.94 | 25 | 0.2556 | 96.362 | 372 | 30.1-31.3 | 192 | 100% |
| <i>psbA-trnH</i> | 0.792 | 14 | 0.00250 | 4.100 | 22 | 29.2-32.5 | 70 | 100% |
| <i>trnL-trnF</i> | 0.914 | 24 | 0.00883 | 6.137 | 52 | 31.9-33.0 | 14 | 100% |
| <i>rbcL</i> | 0.814 | 16 | 0.00376 | 2.273 | 22 | 43.8-44.1 | 10 | 100% |
| <i>trnS-trnG</i> | 0.879 | 20 | 0.02245 | 8.036 | 57 | 20.7-22.4 | 20 | 97.9% |
| <i>trnF-ndhJ</i> | 0.965 | 23 | 0.01652 | 8.295 | 49 | 27.6-29.1 | 27 | 93.7% |
| <i>trnP-psaJ-rpl33</i> | 0.880 | 19 | 0.10146 | 34.902 | 229 | 28.0-30.6 | 195 | 100% |
| <i>psbB-psbH</i> | 0.926 | 21 | 0.03035 | 13.961 | 268 | 34.8-36.5 | 107 | 100% |
| <i>ycf1b</i> | 0.952 | 26 | 0.30907 | 80.975 | 270 | 27.6-32.0 | 162 | 100% |
| <i>ITS</i> | 0.995 | 40 | 0.05201 | 31.776 | 244 | 58.3-64.0 | 105 | 91.6% |

Note: haplotype diversity (Hd), number of haplotypes (H), nucleotide diversity (Pi), average number of nucleotide differences (k), total number of mutations (Eta).

Table 4 Recognition success index based on the program TaxonDNA

| Region | Best match | | | Best Close match | | | All species barcodes | | |
|---|------------|--------------|--------------|------------------|--------------|--------------|----------------------|--------------|--------------|
| | Correct(%) | Ambiguous(%) | Incorrect(%) | Correct(%) | Ambiguous(%) | Incorrect(%) | Correct(%) | Ambiguous(%) | Incorrect(%) |
| <i>ITS+trnP-psaJ-rpl33+psbB-psbH</i> | 12.33 | 12.5 | 79.16 | 12.33 | 10.41 | 79.16 | 8.08 | 95.83 | 0 |
| <i>ITS+trnP-psaJ-rpl33</i> | 10.41 | 10.41 | 79.16 | 10.41 | 10.41 | 75 | 4.16 | 89.58 | 2.08 |
| <i>ITS</i> | 9.09 | 40.9 | 50 | 9.09 | 31.81 | 34.08 | 0 | 75 | 0 |
| <i>ITS+psbB-psbH</i> | 8.33 | 37.5 | 54.16 | 8.33 | 37.5 | 52.08 | 2.08 | 93.75 | 2.08 |
| <i>trnP-psaJ-rpl33+matK</i> | 8.33 | 27.08 | 64.58 | 8.33 | 27.08 | 60.41 | 0 | 95.83 | 0 |
| <i>ITS+psbB-psbH+ycf1b</i> | 8.33 | 8.33 | 83.33 | 8.33 | 6.25 | 83.33 | 4.16 | 93.75 | 0 |
| <i>ITS+trnP-psaJ-rpl34+psbB-psbH+ycf1b</i> | 8.33 | 2.08 | 89.58 | 8.33 | 0 | 89.58 | 4.16 | 91.66 | 2.08 |
| <i>trnP-psaJ-rpl33+psbB-psbH+ycf1b+matK</i> | 8.33 | 2.08 | 89.58 | 8.33 | 0 | 89.58 | 4.16 | 91.66 | 2.08 |
| <i>ITS+ycf1b</i> | 8.16 | 28.57 | 63.26 | 8.16 | 26.52 | 59.18 | 8.16 | 85.71 | 0 |
| <i>ITS+matK</i> | 6.25 | 22.91 | 70.83 | 6.25 | 20.83 | 70.83 | 0 | 97.91 | 0 |
| <i>ycf1b</i> | 6.25 | 66.66 | 27.08 | 6.25 | 64.58 | 27.08 | 6.25 | 79.16 | 12.5 |
| <i>matK</i> | 6.25 | 50 | 43.75 | 6.25 | 47.91 | 43.75 | 0 | 97.91 | 0 |
| <i>trnP-psaJ-rpl33+ycf1b</i> | 6.25 | 45.83 | 47.91 | 6.25 | 43.75 | 47.91 | 6.25 | 87.5 | 4.16 |
| <i>psbB-psbH+ycf1b</i> | 6.25 | 45.83 | 47.91 | 6.25 | 43.75 | 47.91 | 2.08 | 93.75 | 2.08 |
| <i>psbB-psbH+matK</i> | 6.25 | 37.5 | 56.25 | 6.25 | 37.5 | 52.08 | 0 | 95.83 | 0 |
| <i>ITS+psbB-psbH+ycf1b+matK</i> | 6.25 | 6.25 | 87.5 | 6.25 | 6.25 | 81.25 | 0 | 93.75 | 0 |
| <i>ITS+ycf1b+matK</i> | 6.12 | 10.2 | 83.67 | 6.12 | 10.2 | 79.59 | 0 | 95.91 | 0 |
| <i>ITS+trnP-psaJ-rpl34+psbB-psbH+ycf1b+matK</i> | 6.12 | 4.08 | 89.79 | 6.12 | 4.08 | 87.75 | 0 | 95.91 | 2.04 |
| <i>ycf1b+matK</i> | 4.16 | 35.41 | 60.41 | 4.16 | 35.41 | 54.16 | 0 | 91.66 | 3.08 |
| <i>trnP-psaJ-rpl33</i> | 4.16 | 56.25 | 39.58 | 4.16 | 56.25 | 39.58 | 8.33 | 83.33 | 8.33 |
| <i>trnP-psaJ-rpl33+psbB-psbH</i> | 4.16 | 58.33 | 37.5 | 4.16 | 56.25 | 37.5 | 2.08 | 93.75 | 2.08 |
| <i>psbB-psbH+ycf1b+matK</i> | 4.16 | 18.75 | 77.08 | 4.16 | 18.75 | 70.83 | 0 | 91.66 | 2.08 |
| <i>trnP-psaJ-rpl33+psbB-psbH+ycf1b</i> | 0 | 39.58 | 60.41 | 0 | 37.5 | 60.41 | 0 | 95.83 | 2.08 |
| <i>psbB-psbH</i> | 0 | 81.25 | 18.75 | 0 | 81.25 | 16.66 | 0 | 95.83 | 2.08 |

to 34 species and 11 varieties (45 germplasm resources) of 6 sections were selected. The 6 sections include Sect. *Martagon*, Sect. *Archelirion*, Sect. *Sinomartagon*, Sect. *Leucolirion*, Sect. *Pseudolirium*, and Sect. *Liriotypus*. The section and quantity of each sample were listed in Table 1. Fresh leaves were sampled and stored at -80°C until DNA extraction. Total genomic DNA was extracted using a plant genome extraction kit (Tiangen, Beijing, China). Three samples of *Cardiocrinum giganteum*, *Nomocharis pardanthina*, and *Fritillaria karelinii* were used as outgroups.

Primer design and PCR amplification

The software Geneious.10 was used to design 22 pairs of primers (attached file 1: Table S1) [54]. Tissue samples were taken and the products were amplified by Sanger sequencing. The primer fragments with low polymorphism sites and failed sequencing were also removed to determine whether the primer was suitable for *Lilium*. PCR was performed in 25 µL volume consisting of 10 ng of genomic DNA, 0.2 µL of LA Taq (Takara), 2.0 µL of dNTP, 2.5 µL of 10×buffer, 1 µL of upstream and downstream primers (10 µmol/L stock) and ddH₂O

supplemented to 25 µL. Amplification was performed in Veriti Thermal Cycler (Applied Biosystems) with the following program: 3 minutes at 95°C for initial denaturation; denaturation 95°C, 30 seconds; 35 cycles consisting, annealing temperature 50°C - 56°C, 45 seconds, extension temperature 72°C, 1 minute, final 7-minute extension at 72°C, 4°C low temperature save. At least 10 replications of each species were sequenced in both sense and antisense directions. PCR products were separated by electrophoresis on 1.2 % Agarose Gel and visualized with GelRed stain (Biotium, USA).

Sequence alignment and data analysis

To modify heterozygous loci including W, R, Y, S, K, and M, the sequencing results which referred to the Chromas sequence peak map were performed by BioEdit v.7 [55]. Proofreading sequences were sequenced using Geneious's MAFFT v7 (<https://doi.org/10.1093/nar/gkf436>) [56] to quantify sequence length and base composition. Aligned sequences were counted using MEGA v 6.0 for the sequence of variant information [57]. The PHASE operation was performed with DNA SP v5.1 [58] (run-length = 10 000 iterations, burn-in = 100, thinning interval = 10) to calculate the total number of nuclear mutations (Eta), haplotype diversity (Hd), nucleotide diversity information (Pi), average nucleotide difference (k), and the number of haplotypes (H). Pi was the detection of the most essential genetic differences between different individuals, and it represented the existence of different bases between species (sliding window=800 bp, step=200 bp). H refers to the number of different haplotypes contained in all the sequences under study. Hd refers to the frequency of randomly selecting two different haplotypes from a sample. The population with high Hd indicated rich genetic resources. Eta was the site at which a mutation occurs on a base of a sequence. K was the total nucleotide variation /the number of individual samples. Based on similarity obtained from Taxon DNA software, the individual-level discrimination rates of all possible single and combination markers were tested under the K2P-corrected distance model [59]. Taxon DNA with 'pairwise summary function' was used to estimate the barcoding gap. To precisely analyze *Lilium* species, each barcode candidate was measured for appropriate identification proportion by 'best match', 'best close match', and 'all species barcode' functions of Taxon DNA [53]. We evaluated the origin of monophyletic by tree-based analysis to access the effectiveness of marker discriminatory performance. Phylogenetic analyses were performed using maximum likelihood (ML) and Bayesian (BI). ML analyses were performed

by RAXML-HPC BlackBox v.8.1.24 [60] at the CIPRES Science Gateway website [61] (<http://www.phylo.org>). For ML phylogenetic tree analyses, the best-fit models and general time reversible (GTR) + G were used with 1 000 bootstrap replicates. BI was performed with MrBayes 3.2 [62]. The BI model is constructed, and ngen is set to 1 000 000 burnin = ngen*0.25/ Sample-freq, Lset nst = 6 rates = invgamma, Prset statefreqpr = dirichlet (1, 1, 1, 1), graphical visualization of STRU CTURE results using Clumpak. Phylogenetic trees were visualized using Tree view. The best scoring tree was visualized with FigTree 1.3.1 (<http://tree.bio.ed.ac.uk/>).

Abbreviations

ITS: Internal Transcribed Spacer; *rbcl*: ribulose 1,5-bisphosphate carboxylase; *matK*: megakaryocyte-associated tyrosine kinase; *psbA-trnH*: *Acer distylum* *psbA*-*tRNA*-His; *trnL-trnF*: *tRNA*-Leu-*tRNA*-Phe; *trnS-trnG*: *tRNA*-Ser-*tRNA*-Gly; *trnF-ndhJ*: *tRNA*-Phe-*ndhJ*; PCR: Polymerase Chain Reaction; cp genome: chloroplast genome; RAPD: Random Amplified Polymorphic DNA; ISSR: Inter-simple sequence repeat; Pi: the nucleotide polymorphism information; Eta: the total number of mutations; Hd: haplotype diversity; K: average nucleotide difference; K2P: Kimura-2-Parameter; ML: maximum likelihood; BI: Bayesian.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12870-021-03229-6>.

Table S1. Hypervariable region primer design. 22 pairs of primers were designed using the software Geneious 1.10.

Table S2. Information about tested *Lilium* species in this study.

Acknowledgments

We thank Dr. Carina Carianopol (Department of Biological Sciences, University of Toronto) for editing this manuscript.

Authors' contributions

YXL, JW, and YPD conceived the experiments. MFZ, XC, YH, WQP, and YX collected the samples. XQC and JLG conducted the experiments. YXL, XHZ, and JW analyzed the results, and YXL was a major contributor to manuscript writing. All authors read and approved the final manuscript.

Funding

This work was supported by the Natural Science Foundation of China (Grant No.

31601781 and 32171864 to Y. D.; 3217180532 to J. W.), the special projects for capacity-building in scientific and technological innovation of Beijing Academy of Agriculture and Forestry (KJCX20200103 to M. Z.), Beijing Natural Science Foundation (6212012 to J. W.), by the National Key Research and Development Program of China (2020YFD1000400 to Xiuhai Zhang), and the Youth Fund Project of Beijing Academy of Agriculture and Forestry (QNJ201910 to M. Z.). There is no role of the funding body in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

The chloroplast genome sequences of 12 *Lilium* referred in this work were downloaded from the GenBank of NCBI (<https://www.ncbi.nlm.nih.gov/>). The sequences were proofreading using Geneious's MAFFT v7 (<http://mafft.cbrc.jp/alignment/server/>). All the DNA barcoding-Seq raw data are available at NCBI, and accession numbers were listed in Table S2. The PHASE operation was performed with DNA SP v5.1 to calculate the index. The ML Tree was performed by RAXML-HPC BlackBox v.8.1.24 at the CIPRES Science Gateway

website (<http://www.phylo.org>). The BI Tree was performed with MrBayes 3.2. The tree was visualized with FigTree 1.3.1 (<http://tree.bio.ed.ac.uk/>). Plant material was collected from the National Lily Germplasm Bank at Beijing Academy of Agricultural and Forestry Sciences.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China.

²Beijing Key Laboratory of Development and Quality Control of Ornamental Crops, Department of Ornamental Horticulture and Landscape Architecture, China Agricultural University, Beijing 100193, China. ³School of Landscape Architecture, Beijing Forestry University, Beijing 100083, China.

Received: 19 September 2020 Accepted: 23 September 2021

Published online: 13 October 2021

References

- Macrae EA. Lilies: a guide for growers and collectors; 1998.
- Du YP, He HB, Wang ZX, Wei C, Li S, Jia GX. Investigation and evaluation of the genus *Lilium* resources native to China. *Gen Res Crop Evol*. 2014;61(2):395–412.
- Baranova MV. A synopsis of the system of the genus *Lilium* (Liliaceae); 1988.
- Du YP, He HB, Wang ZX, Li S, Wei C, Yuan XN, et al. Molecular phylogeny and genetic variation in the genus *Lilium* native to China based on the internal transcribed spacer sequences of nuclear ribosomal DNA. *Plant Res*. 2014;127(2):249–63.
- Jong PCD. Some notes on the evolution of lilies; 1974.
- Zhao X, Chen X, Li D, Liu K. Resources and research situation of the genus *Lilium* in China. *Acta Horticulturae*. 1996;414:59–68.
- Comber HF. A new classification of the genus *Lilium*. *Lily Year Book of RHS*; 1949.
- Du F, Jiang J, Jia H, Zhao XY, Wang WH, Gao QK, et al. Selection of generally applicable SSR markers for evaluation of genetic diversity and identity in *Lilium*. *Biochem Syst Ecol*. 2015;61:278–85.
- Munafò JP Jr, Gianfagna TJ. Chemistry and biological activity of steroidal glycosides from the *Lilium* genus. *Nat Product Reports*. 2015;32(3):454–77.
- Shin KS, Chakrabarty D, Paek KY. Sprouting rate, change of carbohydrate contents and related enzymes during cold treatment of lily bulblets regenerated in vitro. *Scientia Horticulturae*. 2002;96(1–4):195–204.
- Jin L, Zhang Y, Yan L, Guo Y, Niu L. Phenolic Compounds and Antioxidant Activity of Bulb Extracts of Six *Lilium* Species Native to China. *Molecules*. 2012;17(12):9361.
- Nishikawa T, Okazaki K, Nagamine T. Phylogenetic Relationships among *Lilium auratum* Lindley, *L. auratum* var. *platyphyllum* Baker and *L. rubellum* Baker based on three spacer regions in chloroplast DNA. *Breeding Sci*. 2002;52(3):207–13.
- Nair PKK. Plant taxonomy. *CUR SCI*. 2004;86(5):665–7.
- Barbi G, Singh BB. DNA barcoding of the genus *Nepenthes* (Pitcher plant): a preliminary assessment towards its identification. *BMC Plant Biol*. 2018;18(1):153.
- Liu YX, Huang J, Moe TS, Khan MS, Xue J, Zhang XF, et al. The complete chloroplast genome sequence of *Lilium speciosum* var. *gloriosoides*, an important breeding parent. *Mitochondrial DNA Part B*. 2020;5(1):71–2.
- Iguchi T, Yokosuka A, Kuroda M, Takeya M, Hagiya M, Mimaki Y. Steroidal glycosides from the bulbs of *Lilium speciosum*. *Phytochem Lett*. 2020;37:21–8.
- Zhang Q, Bi Y, Zhang MF, Chen XQ, Yang FP, Xue J, et al. The complete chloroplast genome of *Lilium taliense*, an endangered species endemic to China. *Conservation Gen Res*. 2017;9(2):201–3.
- Wang S, Xie Y. China species red list; 2004.
- Du YP, Bi Y, Yang FP, Zhang MF, Chen XQ, Xue J, et al. Complete chloroplast genome sequences of *Lilium*: insights into evolutionary dynamics and phylogenetic analyses. *Sci Reports*. 2017;7(1).
- Bi Y, Du YP, Chen XQ, Yang FP, Xue J, Zhang XH, et al. The complete chloroplast genome sequence of *Lilium fargesii* (Lilium, Liliaceae). *Conservation Gen Resour*. 2016;8(4):419–22.
- Vere ND, Rich TCG, Trinder SA, Long C. DNA Barcoding for Plants. *Methods Mole Bio*. 2015;1245:101–18.
- Yang XW, Henry RJ, Rossetto W, et al. Plant DNA barcoding: from gene to genome. *Biol Rev*. 2015;90(1):157–66.
- Ashfaq M, Akhtar S, Khan AM, Adamowicz SJ, Hebert PDN. DNA barcode analysis of butterfly species from Pakistan points towards regional endemism. *Mole Eco Resour*. 2013;13(5):832–43.
- Vogler AP, Monaghan MT. Recent advances in DNA taxonomy. *J Zoological System Evol Res*. 2010;45(1):1–10.
- Dong WP, Xu C, Li C, Sun J, Zuo Y, Shi S, et al. *ycf1*, the most promising plastid DNA barcode of land plants. *Sci Reports*. 2015;5:8348.
- Hollingsworth PM, Forrest LL, Spouge JL, et al. A DNA barcode for land plants. *Proceed Nat Acad Sci*. 2009;106(31):12794–7.
- Vere ND, Rich T, Ford CR, Trinder SA, Long C, Moore CW, et al. DNA barcoding the native flowering plants and conifers of Wales. *PLOS ONE*. 2012;7(6):e37945.
- Robert V, Cardinali G, Stielow B, Vu TD, Schoch C. Fungal DNA Barcoding. 2015.
- Giudicelli GC, Mäder G, Freitas LBD. Efficiency of ITS Sequences for DNA Barcoding in Passiflora (Passifloraceae). *Int J Mole Sci*. 2015;16(4):7289–303.
- Chen S, Hui Y, Han J, Liu C, Song J, Shi L, et al. Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLOS ONE*. 2010;5(1):e8613.
- Chase M, Fay W, Michael F. Barcoding of plants and fungi. *Science*. 2009.
- Wang W, Wu Y, Yan Y, Ermakova M, Kerstetter R, Messing J. DNA barcoding of the Lemnaceae, a family of aquatic monocots. *BMC Plant Biol*. 2010;10(1):205.
- Steinke D, de Vere N, Rich TCG, Ford CR, Trinder SA, Long C, et al. DNA barcoding the native flowering plants and conifers of Wales. *PLOS ONE*. 2012;7(6):e37945.
- Small RL, Cronn RC, Wendel JF. Use of nuclear genes for phylogeny reconstruction in plants. *Australian Syst Bot*. 2004;17.
- Brozynska M, Furtado A, Henry RJ. Genomics of crop wild relatives: expanding the gene pool for crop improvement. *Plant Biotechnol J*. 2016;1070–85.
- Huang H, Shi C, Liu Y, Mao SY, Gao LZ. Thirteen *Camellia* chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships. *BMC Evol Biol*. 2014;14(1):151.
- Yang JB, Yang SX, Li HT, Jing Y, Li DZ, Turgay U. Comparative chloroplast genomes of *Camellia* species. *Plos One*. 2013;8(8):e73053.
- Lei W, Ni D, Wang Y, Shao J, Wang X, Yang D, et al. Intraspecific and heteroplasmic variations, gene losses and inversions in the chloroplast genome of *Astragalus membranaceus*. *Entific Reports*. 2016;6:21669.
- Lakhani HN, Patel SV, Bodar NP, Golakiya BA. RAPD analysis of genetic diversity of castor bean (*Ricinus communis* L.). *Int J Curr Microbiol Appl Sci*. 2015;4(1):696–703.
- Lu X, Yawo M, Nevame A, Gilbert N, Chofong M. Identification of high-efficiency SSR markers for assessing watermelon genetic purity. *Journal of Genetics*. 2018;97(5):1295–306.
- Frederic A, Olivier D, Brigitte S, et al. DNA barcode analysis: a comparison of phylogenetic and statistical classification methods. *BMC Bioinform*. 2009.
- Liu M, Li XW, Liao BS, Luo L, Ren YY. Species identification of poisonous medicinal plant using DNA barcoding. *Chinese J Nat Med*. 2019;17(8):585–90.
- Li HQ, Chen JY, Wang S, Xiong SZ. Evaluation of six candidate DNA barcoding loci in *Ficus* (Moraceae) of China. *Mol Ecol Resour*. 2012;12(5):783–90.

44. Kim HM, Oh SH, Bhandari GS, Kim CS, Park CW. DNA barcoding of Orchidaceae in Korea. *Mol Ecol Resour*. 2014;14(3):499–507.
45. Evans J, Sheneman L, Foster J. Relaxed neighbor joining: a fast distance-based phylogenetic tree construction method. *J Mole Evol*. 2006;62(6):785–92.
46. Liu HY, Li J, Xie DF, He XJ, Zhou SD. The complete chloroplast genome of *Nomocharis pardanthina*. *Mitochondrial DNA Part B*. 2018;3(1):103–4.
47. Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH. Use of DNA barcodes to identify flowering plants. *Proceed Nat Acad Sci*. 2005;102(23):8369–74.
48. Yan LJ, Liu J, Ller MM, Zhang L, Zhang XM. DNA barcoding of *Rhododendron* (Ericaceae), the largest Chinese plant genus in biodiversity hotspots of the Himalaya–Hengduan Mountains. *Mol Ecol Resour*. 2015;15(4):932–44.
49. Vázquez ML. Molecular evolution of the internal transcribed spacers in red oaks (*Quercus* sect. *Lobatae*). *Comput Biol Chem*. 2019;83:107117.
50. Lahaye R, Bank MVD, Bogarin D, Warner J, Savolainen V. DNA barcoding the floras of biodiversity hotspots. *Proceed Nat Acad Sci*. 2008;105(8):2923–8.
51. Hollingsworth PM, Graham SW, Little DP. Choosing and Using a Plant DNA Barcode. *PLOS ONE*. 2011;6(5).
52. Li DZ, Gao LM, Li HT, Wang H, Ge XJ, Liu JQ, et al. Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proceed Nat Acad Sci*. 2011.
53. Denton JSS, Goolsby EW. Measuring inferential importance of taxa using taxon influence indices. *Ecol Evol*. 2018;8(9):4484–94.
54. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinform*. 2012;28(12):1647–9.
55. Hall TA. BioEdit: A User-Friendly Biological Sequence Alignment Editor and Analysis Program for Windows 95/98/NT. *Nucleic Acids Symposium Series*. 1999;41(41):95–8.
56. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. 2002;30(14):3059–66.
57. Sudhir K, Glen S, Koichiro T. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mole Biol Evol*. 2016;7(1):1870.
58. Librado R. DNAsp v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics (Oxford, England)*. 2009;25.11:1451–2.
59. Meyer CP, Paulay G. DNA Barcoding: Error Rates Based on Comprehensive Sampling. *PLOS Biol*. 2005;3(12).
60. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014;30(9).
61. Miller MA, Pfeiffer W, Schwartz T. Creating the CIPRES Science Gateway for Inference of Large Phylogenetic Trees. *Gateway Computing Environments Workshop (GCE)*. IEEE Comput Soc. 2010.
62. Huelsenbeck JP. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *System Biol*. 2012;61(3):539–42.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

